Permuted Pattern Matchings on Multi-Track Strings

氏名（本籍地）	桂（かつら）　敬史（たかし）
学位の種類	博　士（情報科学）
学位記番号	情　博　第　５９３号
学位授与年月日	平成 27 年　3 月 25 日
学位授与の要件	学位規則第 4 条第 1 項該当
研究科、専攻	東北大学大学院情報科学研究科（博士課程）システム情報科学専攻
学位論文題目	Permuted Pattern Matchings on Multi-track Strings （マルチトラック文字列上の順列パターン照合）
論文審査委員	（主査）東北大学 教授　篠原　歩 東北大学 教授　徳山　豪 東北大学 教授　周　暁

論 文 内 容 の 要 旨

第 1 章　序論

文字列のパターン照合は情報検索における最も基本的な技術であり，文字列検索をはじめ，ゲノム情報処理やデータ圧縮など様々な応用をもつ．数値列で表される系列データも文字列への変換処理を行うことで，文字列のパターン照合を適用することができる．近年，センサ技術などの発達により，複数の系列が同時に観測される例が増えている．今や身近な存在となったスマートフォンにも 3 軸加速度センサをはじめ様々なセンサが取り付けられおり，複数のセンサデータを容易に取得できる．このような多系列データ上で高度な解析を行う際には，系列の組合せを考慮した照合を行うことが重要になると考えられる．例えば，スマートフォンが斜めに動いたことを検知するには3軸の加速度センサのうち2つの軸を選ぶ組合せに注目して照合を行う必要がある．しかし，従来のパターン照合アルゴリズムは単一文字列間の照合を念頭に置いているため，多系列データにそのまま適用することは難しい．

本研究の目的は，系列間の組合せを考慮に入れた多系列データの高速な照合手法の開発である．本研究では文字列の多重集合をマルチトラック文字列，あるいは単にマルチトラックと呼び，多系列データのモデルとして扱う．マルチトラックに対する効率的な照合アルゴリズムを開発することで，多系列データ処理の基盤技術を開発する．具体的には，長さnのN個の文字列からなるマルチトラックテキストから，長さmのM個の文字列からなるマルチトラックパターンが出現する位置を探すパターン照合問題に取り組む．このような照合は既存研究にはなかった枠組みであり，文字列の順列組合せを考慮した照合として捉えることができることから順列パターン照合問題と呼ぶ．

第 2 章　準備

第 2 章では，文字列に関する諸定義や，先行研究におけるデータ構造・アルゴリズムを説明し，提案手法の導入の準備を行う．具体的には，文字列，トライ木，シーケンスハッシュ木，根付き木上のクエリ，索引構造，スペクトルブルームフィルタ，ハッシュ関数についての説明を行う．

第 3 章　マルチトラック

第 3 章では，本論文で扱うマルチトラック文字列と順列パターン照合問題を定義する．

文字の有限集合をΣで表し，アルファベットと呼ぶ．Σ^*の要素を文字列と呼ぶ．文字列x, yについて，xとyの連結をxyで表す．文字列tについて，tの長さを$|t|$で表し，$1 \leq i \leq j \leq |t|$に対して，$t[i:j]$を位置iから位置jまでのtの部分文字列とする．

Σ上の長さnの文字列のN項組をマルチトラック文字，または単にマルチトラックと呼ぶ．マルチトラック$X = (x_1, x_2, \ldots, x_N)$，$Y = (y_1, y_2, \ldots, y_N)$について，$X$と$Y$の連結を$XY = (x_1y_1, x_2y_2, \ldots, x_Ny_N)$とする．マルチトラック$T = (t_1, t_2, \ldots, t_N)$に対し，$T$の$i$番目の要素$t_i$を$i$番目のトラックと呼ぶ．$T$の長さを$|T|_{len} = |t_1| = |t_2| = \cdots = |t_N| = n$で表し，$T$のトラック数を$|T|_{num} = N$で表す．マルチトラック$T = XYZ$に対し，$X$, Y, ZをそれぞれTの接頭辞，部分文字列，接尾辞と呼ぶ．$T[i]$は$(t_1[i], t_2[i], \ldots, t_N[i])$を表す．$1 \leq i \leq j \leq n$に対し，位置$i$から始まり位置$j$で

終わるTの部分文字列を$T[i:j] = (t_1[i:j], t_2[i:j], \dots, t_N[i:j])$で表す.

マルチトラック$X = (x_1, x_2, \dots, x_N)$を考える.$r = (r_1, r_2, \dots, r_K)$を$(1, \dots, N)$の順列とする.ここで,$1 \leq K \leq N$である.$r$によって特定される$X$の順列マルチトラックとはマルチトラック$(x_{r_1}, x_{r_2}, \dots, x_{r_K})$である.

$|X|_{len} = |Y|_{len}$,$|X|_{num} \leq |Y|_{num}$を満たす任意のマルチトラックX,Yに対し,Yの順列マルチトラックY'が存在して$X = Y'$を満たすとき,XはYと順列一致するという.特に,$|X|_{num} = |Y|_{num}$の場合,XはYと全順列一致するという.$|X|_{num} < |Y|_{num}$の場合,XはYと部分順列一致するという.

本論文で取り組む問題は形式的に以下のように定義される.長さnのマルチトラックテキスト$T = (t_1, \dots, t_N)$と長さmのマルチトラックパターン$P = (p_1, \dots, p_M)$が与えられたとき,PがTと順列一致する位置iを出力する.$N = M$のとき,この問題を全順列パターン照合問題と呼び,$N > M$のとき,部分順列パターン照合問題と呼ぶ.

第 4 章　順列パターン照合アルゴリズム

本章では,マルチトラック文字列上の順列パターン照合を高速に計算するアルゴリズムを,(1) 前処理を行わない,(2) Pのみ前処理,(3) Tのみ前処理の 3 つの場合それぞれについて提案する.

まず,4.1 節では,TとPのいずれも前処理しない場合について,一般化接尾辞配列を用いたアルゴリズムを示す.本手法では,マルチトラックに含まれるトラックを辞書式順序でソートすることで順列一致が効率的に判定可能であることに注目して,テキストとパターンに含まれるすべてのトラックに対する一般化接尾辞配列を用いる.接尾辞配列とは,文字列のすべての接尾辞を辞書式順序で小さい順に並べた配列である.文字列tの接尾辞配列は$O(|t|)$時間で構築可能であることが知られている.また,接尾辞配列を利用したアルゴリズムとして,tの 2 つの接尾辞の最長共通接頭辞長を定数時間で求める最長共通拡張クエリが知られている.これらを利用することで順列パターン照合問題が$O(nN)$時間で計算可能であることを示す.

次に,4.2 節では,Pのみ前処理可能な場合について,Aho-Cosrasick (AC)オートマトンを用いたアルゴリズムを提案する.AC オートマトンはテキスト文字列に対して複数のパターン文字列を照合するための手法である.パターンを受理するオートマトンに対して,テキストの各文字を使って状態遷移を行うことで照合を行う.順列パターン照合においては,AC オートマトンの各受理状態に多重度をもたせ,テキストTの各トラックに対応したN個のポインタを用いて,各々独立に状態遷移を行うことで順列一致を判定する.上記の拡張を行っても,AC オートマトンがPのサイズに線形な時間で構築可能であり,AC オートマトンをあらかじめ構築しておくことで,テキストTのサイズに線形な$O(nN)$時間で照合可能であることを示す.

最後に,4.3 節では,Tのみ前処理可能な場合に関して,全順列パターン照合を効率的に行うためのマルチトラックの索引構造の提案を行う.索引構造とは巨大なテキストから特定のパターンを高速に検索するための辞書である.上述した接尾辞配列をはじめ,さまざまな文字列の索引構造が知られている.本論文では,文字列の索引構造である接尾辞木,および,ポジションヒープを基にした,3 つのマルチトラックの索引構造を提案し,これらの構造がテキストTの大きさに線形な$O(nN)$時間で構築可能であることを示す.図 1 に提案するデータ構造の例を示す.また,接尾辞木を基にしたマルチトラック接尾辞木を用いることで,全順列パターン照合が$O(mN)$時間で計算可能であることを示す.一方,ポジションヒープを基にした縮約マルチトラックポジションヒープは,照合時間が$O(m^2N^2)$時間となるものの,必要領域が高々$O(n)$の省メモリなデータ構造であることを示す.

第 5 章　近似順列パターン照合アルゴリズム

順列パターン照合問題におけるマルチトラック文字列を数値列へと拡張し,マルチトラック数値列上の距離を定義することにより,より実用的な近似照合へと拡張できる.第 5 章では,このような近似順列パターン照合にも応用可能なデータ構造を提案する.

順列パターン照合問題を文字列の多重集合の包含判定と捉えたとき,スペクトルブルームフィルタ (SBF)と呼ばれるデータ構造を用いることで高速に判定を行うことが可能である.SBF は多重集合の要素に対するハッシュ値の出現数を保持した配列であり,2 つの多重集合に関してそれ

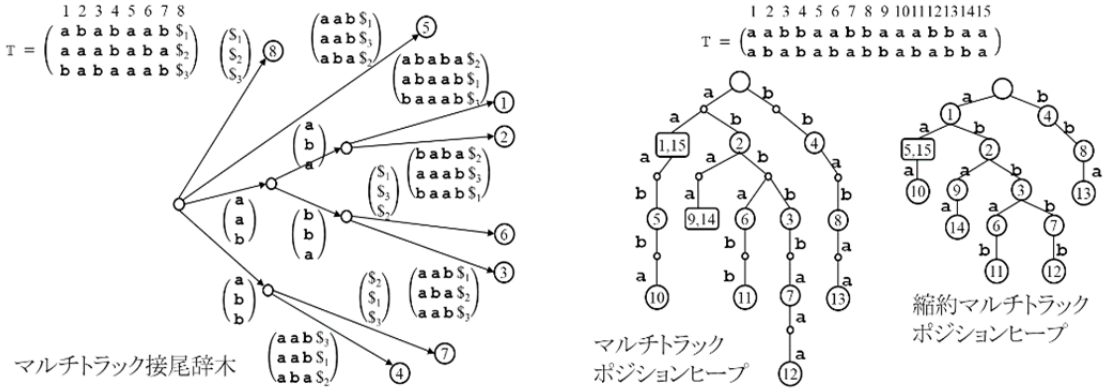

図 1　マルチトラックの索引構造の例

表 1　順列パターン照合に対するデータ構造

データ構造	必要領域	構築時間	照合時間
一般化接尾辞配列	$O(nN)$	$O(nN)$	$O(nN)$
AC オートマトン	$O(mM)$	$O(mM)$	$O(nN)$
マルチトラック接尾辞木	$O(nN)$	$O(nN)$	$O(mN)$
マルチトラックポジションヒープ	$O(nN)$	$O(nN)$	$O(m^2N)$
縮約マルチトラックポジションヒープ	$O(n)$	$O(nN)$	$O(m^2N^2)$
FILM 木	$O(n)$	$O(nN)$	$O(mM+n)$

ぞれの SBF の配列の値を比較することで偽陽性のある包含判定が可能である．

本論文で提案するデータ構造 Filtering Multi-set (FILM)木は，各頂点が SBF として表される完全二分木である．FILM 木の各葉はテキストTの長さmの部分文字列に対する SBF であり，各内部頂点は子の SBF を配列値の和をとった SBF である．パターンPに対して同様に SBF を構築し，FILM 木の根から深さ優先で包含判定を行っていくことで，順列パターン照合を$O(mM+n)$時間で計算できる．

SBF 構築に用いるハッシュ関数が文字列に対するものであれば，FILM 木はマルチトラック文字列上の順列パターン照合に対応する．一方，ハッシュ関数が数値列に対するものならば近似照合にも応用可能である．

第 6 章　実験

第 6 章では，第 4 章と第 5 章で提案した手法の性能を計算機実験によって定量的に評価する．実験は大きく 4 つ行い，それぞれ (1) 順列パターン照合アルゴリズム（一般化接尾辞配列，AC オートマトン，マルチトラック接尾辞木）の比較，(2) マルチトラックに対する索引構造の比較，(3) FILM 木の性能の検証，(4) 索引構造と FILM 木の比較を行った．

第 7 章　結論

第 7 章では，本論文の結論と今後の課題を述べる．

本論文では系列間の組合せを考慮に入れた多系列データの高速な照合手法の開発を目的として，マルチトラック文字列上の順列パターン照合問題を新たに定義し，これを高速に解くアルゴリズムの提案を行った．本論文の主結果を表 1 に示す．

今後の課題として，部分順列パターン照合も可能な索引構造や，各索引構造のオンライン構築アルゴリズムの提案が挙げられる．また，FILM 木による照合を任意のパターン長に拡張した場合の解析も課題の 1 つである．

論文審査結果の要旨

文字列パターン照合は，情報検索における最も基本的な問題であり，ウェブ検索やテキストマイニング，遺伝子情報処理など，大量のデータを効率よく処理するための基盤技術として，理論と応用の両面からさまざまな研究が行われてきた．近年，センサ技術の発達を背景に，多数のセンサによって同時に観測されたデータが大量に取得できるようになり，その利活用への期待が高まっている．例えば，多数の自動車のプローブデータから得られる速度と位置情報を収集し，統合処理することによって，交通量や路面状況の推定，事故発生の検出，渋滞の予測などが行われるようになってきている．このような多系列データを解析する上では，個々の系列データを個別に処理するだけではなく，系列の組合せに注目したパターン照合が有用である．従来の単純なパターン照合は単一系列での照合を想定したものであり，多系列間の高度な照合を効率よく行うことは困難であった．

著者は，多系列データ上の系列間の組合せを考慮した照合手法の開発を主目的として，多系列データをマルチトラック文字列としてモデル化し，マルチトラック文字列上の順列パターン照合問題を新たに提案した．さらに，文字列学の観点から，順列パターン照合問題を高速に解くアルゴリズムの開発に取り組んできた．本論文はその成果をまとめたもので，全編 7 章からなる．

第 1 章は序論である．

第 2 章では，準備として論文中で用いる用語や記号について基本的な定義を与えている．

第 3 章では，本論文で提案するマルチトラック文字列と順列パターン照合問題を定式化すると共に，基本的なアルゴリズムの説明を行っている．

第 4 章では，マルチトラック文字列上の厳密な順列パターン照合を高速に行うためのアルゴリズムを，(1) 前処理を行わない，(2) パターンのみ前処理，(3) テキストのみ前処理，という 3 つの場合それぞれについて提案している．(1) については，テキスト長に線形な時間で問題を解くアルゴリズムを提案している．また，(2)についてはパターンに対して線形時間の前処理を行うことでテキスト長に線形な時間で照合が行えること，(3) についてはテキストに対して線形時間の前処理を行うことでパターン長に線形な時間で照合が行えることを示している．

第 5 章では，部分順列パターン照合において，照合位置を高速にフィルタリングするための確率的データ構造の提案を行っている．提案されたデータ構造は，文字列だけではなく数値列上の近似照合にも適用ができ，幅広い応用が期待される成果である．

第 6 章では，第 4 章と第 5 章で提案した手法の性能を計算機実験によって定量的に評価し，有効性を検証している．

第 7 章は結論であり，論文の成果をまとめると共に，今後の研究課題について述べている．

以上要するに本論文は，マルチトラック文字列上の順列パターン照合問題を新たに定式化し，それを高速に解くためのアルゴリズムとデータ構造について研究したものであり，文字列学や計算量理論を中心にシステム情報科学の発展に寄与するところが少なくない．

よって，本論文は博士（情報科学）の学位論文として合格と認める．